Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Avilash Srivastava: avilashsrivastava@gmail.com

1. Exploratory data analysis – univariate and multivariate analysis.
2. Data Wrangling – checking missing values, outliers, features modification.
3. Fitting Models – splitting the data, applying algorithms, evaluating, model explanation.
4. Presentation, Technical documentation.

Please paste the GitHub Repo link.

Github Link:- https://github.com/avilashsrivastava/Bike-Sharing-Demand-Prediction

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

The contents of the data came from a city called Seoul. It is the capital city of South Korea and has a population of around 9.7 million people. It was the 4th largest metropolitan economy in 2014. It has humid continental climate influenced by monsoons. The data had variables such as date, hour, temperature, humidity, windspeed, visibility, dew point temperature, solar radiation, rainfall, snowfall, seasons, holiday, functioning day and rented bike count.

The problem statement was to build a machine learning model that could predict the rented bikes count required for an hour, given other variables.

The first step in the exercise involved exploratory data analysis where we tried to dig insights from the data in hand. It included univariate and multivariate analysis in which we identified certain trends, relationships, correlation and found out the features who had some impact on our dependent variable.

The second step was to clean the data and perform modifications. We checked for missing values and outliers and removed irrelevant features. We also encoded the categorical variables.

The third step was to try various machine learning algorithms on our splitted and standardized data. We tried 3 different algorithms namely; Linear regression, Randomforest and XGBoost. We did hyperparameter tuning and evaluated the performance of each model using various metrics. The best performance was given by the XGBoost model where the R2_score for training and test set was 0.84 and 0.8 respectively.

Next, we implemented shap techniques to understand the working of our model. The most important features who had a major impact on the model predictions were; hour, temperature, windspeed, solar-radiation, month and seasons. Demand for bikes got higher when the temperature and hour values were more. Demand was high for low values of windspeed and solar radiation. Demand was high during springs and summer and very low during winters.

The model performed good in this case but as the data is time dependent, values of temperature, windspeed, solar radiation etc. will not always be consistent. Therefore, there will be scenarios where the model might not perform well. As Machine learning is an exponentially evolving field, we will have to be prepared for all contingencies and also keep checking our model from time to time.